Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model

Year: 2023 | Citations: 410 | Authors: Douglas B. Johnson, Rachel S Goodman, J. Patrinely, Cosby A Stone, Eli Zimmerman

---

Abstract

Background: Natural language processing models such as ChatGPT can generate text-based content and are poised to become a major information source in medicine and beyond. The accuracy and completeness of ChatGPT for medical queries is not known. Methods: Thirty-three physicians across 17 specialties generated 284 medical questions that they subjectively classified as easy, medium, or hard with either binary (yes/no) or descriptive answers. The physicians then graded ChatGPT-generated answers to these questions for accuracy (6-point Likert scale; range 1 – completely incorrect to 6 – completely correct) and completeness (3-point Likert scale; range 1 – incomplete to 3 - complete plus additional context). Scores were summarized with descriptive statistics and compared using Mann-Whitney U or Kruskal-Wallis testing. Results: Across all questions (n=284), median accuracy score was 5.5 (between almost completely and completely correct) with mean score of 4.8 (between mostly and almost completely correct). Median completeness score was 3 (complete and comprehensive) with mean score of 2.5. For questions rated easy, medium, and hard, median accuracy scores were 6, 5.5, and 5 (mean 5.0, 4.7, and 4.6; p=0.05). Accuracy scores for binary and descriptive questions were similar (median 6 vs. 5; mean 4.9 vs. 4.7; p=0.07). Of 36 questions with scores of 1-2, 34 were re-queried/re-graded 8-17 days later with substantial improvement (median 2 vs. 4; p<0.01). Conclusions: ChatGPT generated largely accurate information to diverse medical queries as judged by academic physician specialists although with important limitations. Further research and model development are needed to correct inaccuracies and for validation.